TITLE:

Pitman's model of random partitions (5th Workshop on Stochastic Numerics)

AUTHOR(S):

Sibuya, Masaaki; Yamato, Hajime

CITATION:

Sibuya, Masaaki ...[et al]. Pitman's model of random partitions (5th Workshop on Stochastic Numerics). 数理解析研究所講究録 2001, 1240: 64-73

ISSUE DATE:

2001-12

URL:

http://hdl.handle.net/2433/41603

RIGHT:

京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

数理解析研究所講究録 1240 巻 2001 年 64-73

Pitman's model of random partitions

Masaaki Sibuya and Hajime Yamato
Takachiho University and Kagoshima University

Random partition of a finite set or number is rudimental in applied probability and statics. The most elementary family of random partitions of a number is Ewens' one-parameter family of random partitions, known as Ewens' sampling formula, which has been developed in the population genetics. See, for example, Ewens (1990) and Johnson et al. (1997).

In a series of papers Pitman enlarged Ewens family to a two-parameter family of random partitions, which will be called Pitman's model, Pitman (1955-1999). In this report the estimation of parameters of Pitman's model is discussed. Further, geneses of random partitions and its statistical applications are reviewed.

1 Random partition of a finite set

1.1 An urn model

Balls $B_1, B_2, \ldots$, are randomly and sequentially put into urns $U_1, U_2, \ldots$ Ball B_1 is put into U_1 with probability 1. If $B_1, \ldots, B_n$ are in $U_1, \ldots, U_k$, in such a way that $c_j > 0$ balls are in U_j, $j = 1, \ldots, k$, $\sum_{j=1}^{k} c_j = n$, ball B_{n+1} is put into

a new urn U_{k+1} with probability $(\theta + k\alpha)/(\theta + n)$,

an old urn U_j with probability $(c_j - \alpha)/(\theta + n)$, $1 \le j \le k$.

The partition of balls into urns is, in terms of the subscript of balls, a partition of the generic set $\mathcal{N}_n = \{1, \ldots, n\}$ into an ordered subsets

$$(a_1, \ldots, a_k), \quad a_i \cap a_j = \emptyset,\ i \neq j\ ;\ \cup_{j=1}^{k} a_j = \mathcal{N}_n.$$

Its probability is, by induction,

$$p((a_1, \ldots, a_k); \theta, \alpha) = \frac{1}{\theta^{[n]}} \prod_{j=1}^{k} (\theta + (j-1)\alpha)(1-\alpha)^{[c_j - 1]}, \tag{1}$$

$$c_j = |a_j| > 0, \quad 1 \le k \le n,$$

where $|a_j|$ denotes the cardinality (number of elements) of a_j, and $x^{[n]} = x\,(x+1) \cdots (x+n-1) = (x+n-1)^{(n)}$. Note that the probability does not depend on the elements of the subsets, nor on the permutation of $\{c_1, \ldots, c_k\}$.

Let $\mathcal{A}_n^o$ denotes the set of all ordered partition of $\mathcal{N}_n$, and call a discrete probability distribution on $\mathcal{A}_n^o$ *random partition.* The random partition (1) is denoted by $\mathcal{P}(\theta, \alpha; \mathcal{A}_n^o)$.

If $\alpha = 0$ everything is simplified, and, for example,

$$p((a_1, \ldots, a_k); \theta, 0) = \frac{\theta^k}{\theta^{[n]}} \prod_{j=1}^{k} (c_j - 1)!, \quad c_j = |a_j| > 0, \quad 1 \le k \le n.$$

1.2 Restriction on the parameters

If $\alpha = 1$, B_n is put into U_n with probability 1, $n = 1, 2, \ldots$ If $\theta = -\alpha$, B_n is put into U_1 with probability 1, $n = 1, 2, \ldots$ These are the degenerating boundaries. The probabilities are nonnegative if only if

$$0 \le \alpha \le 1 \text{ and } -\alpha \le \theta; \quad \text{or} \quad \alpha < 0 \text{ and } \theta = -M\alpha, \ M = 1, 2, \ldots$$

In the last case, B_n, $n = 1, 2, \ldots$, enters B_{M+1} with the probability 0.

1.3 Size index

The number of balls in urns $(c_1, \ldots, c_k)$ is an ordered partition of number n into a sum of positive numbers. Let $\mathcal{C}_n^o$ denote the set of all such partitions.

If the order $(c_1, \ldots, c_k)$ is disregarded, its 'order statistics' is expressed by the set $\{c_1, \ldots, c_k\}$, $c_{(1)} \ge c_{(2)} \ge \cdots \ge c_{(k)}$, or by its *size index*

$$s = (s_1, \ldots, s_n), \quad s_j = |\{i : c_i = j, i = 1, \ldots, k\}| \ge 0, j = 1, \ldots, n;$$
$$\sum_{j=1}^{n} s_j = k \quad \text{and} \quad \sum_{j=1}^{n} j s_j = n,$$

Let $\mathcal{A}_n^u$ denote the set of all unordered partitions of $\mathcal{N}_n$, and the corresponding random partition denoted by $\mathcal{P}(\theta, \alpha; \mathcal{A}_n^u)$ is

$$p(\{a_1, \ldots, a_k\}; \theta, \alpha) = \frac{1}{\theta^{[n]}} \prod_{i=1}^{k} (\theta + (i-1)\alpha) \prod_{j=1}^{n} \left((1-\alpha)^{[j-1]}\right)^{s_j}, \tag{2}$$

where $(s_1, \ldots, s_n)$ is the size index of $(|a_1|, \ldots, |a_k|)$, and $k = \sum_{j=1}^{n} s_j$.

1.4 Types of partitions and random partition of numbers.

In the urn model balls and urns may be distinguishable or undistinguishable. Hence, there are four possible types of the partition.

Two of them are mentioned already, and if both balls are undistinguishable, observable is the partition of a positive number to a sum of positive numbers. See Table 1 for the types of partitions. Each entry is a symbol denoting the set of all possible partitions. The corresponding distribution is denoted using them.

The corresponding probabilities are

Table 1: The Set of All Partitions of Four Types

urns	balls distinguishable	balls undistinguishable
distinguishable	$\mathcal{A}_n^o$	$\mathcal{C}_n^o$
undistinguishable	$\mathcal{A}_n^u$	$\mathcal{C}_n^u$

$\mathcal{P}(\theta,\alpha;\mathcal{A}_n^o)$: (1); $\mathcal{P}(\theta,\alpha;\mathcal{A}_n^u)$: (2), and the others are as follows

$$\mathcal{P}(\theta,\alpha;\mathcal{C}_n^o) : p((c_1,\ldots,c_k);\theta,\alpha)$$
$$= \frac{n!\theta(\theta+\alpha)\cdots(\theta+(k-1)\alpha)}{\theta^{[n]}} \prod_{i=1}^{k} \frac{(1-\alpha)^{[c_i-1]}}{(\sum_{j=i}^{k} c_j)(c_i-1)!}, \tag{3}$$

$$\mathcal{P}(\theta,\alpha;\mathcal{C}_n^u) : p(s;\theta,\alpha)$$
$$= \frac{n!\theta(\theta+\alpha)\cdots(\theta+(k-1)\alpha)}{\theta^{[n]}} \prod_{j=1}^{n} \left(\frac{(1-\alpha)^{[j-1]}}{j!}\right)^{s_j} \frac{1}{s_j!}, \tag{4}$$

$$s=(s_1,\ldots,s_n)\in\mathcal{C}_n^u, \quad k=\sum_{j=1}^{n} s_j.$$

$\mathcal{P}(\theta,0;\mathcal{C}_n^u)$ is known as *Ewens' sampling formula.*

2 Specific properties

2.1 Partition structure

The random partitions $\mathcal{A}_n^o$ and $\mathcal{A}_n^u$ are essentially the same. Any partition of $\mathcal{A}_n^u$ can be ordered by the elements of subsets: The subset (urn) containing 1 is the first subset a_1. The subset containing $\min(\mathcal{N}_n - \cup_{i=1}^{j-1} a_i)$ is the j-th subset a_j. The difference between $\mathcal{A}_n^o$ and $\mathcal{A}_n^u$ is to regard $\mathcal{N}_n$ as a generic finite set or a linearly ordered subset.

Proposition 1. *For any renumbering (permutation) of the elements, $\mathcal{P}(\theta,\alpha;\mathcal{A}_n^o)$ and $\mathcal{P}(\theta,\alpha;\mathcal{A}_n^u)$ are invariant.*

Proposition 2. *From the random partition $\mathcal{P}(\theta,\alpha;\mathcal{A}_n^o)$ sample one element at random with the equal probability $1/n$, and delete it. The result is $\mathcal{P}(\theta,\alpha;\mathcal{A}_{n-1}^o)$ if the elements are renumbered.*

Further, if the selected one belongs to the set a_j delete a_j. The result is $\mathcal{P}(\theta,\alpha;\mathcal{A}_{n-c_j}^o)$, $c_j = |a_j|$.

Proposition 3. *Let $\{j_1,\ldots,j_k\}\in\mathcal{N}_n$, In $\mathcal{P}(\theta,\alpha;\mathcal{A}_n^o)$ the probability that $\{j_1,\ldots,j_k\}$ is the same set. This is equal to the probability that $\{1,\ldots,k\}$ is the same set in $\mathcal{P}(\theta,\alpha;\mathcal{A}_k^o)$,*

$$(1-\alpha)^{[k-1]}/\theta^{[n]}.$$

2.2 Size biased random permutation

The ordering of $\mathcal{P}(\theta, \alpha; \mathcal{A}_n^u)$ into $\mathcal{P}(\theta, \alpha; \mathcal{A}_n^o)$ suggests a similar ordering of $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^u)$ into $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^o)$.

Let $\{c_1, \ldots, c_k\}$ be a partition of n following $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^u)$. The simple random sampling of a ball is now to choose c_{j_1} with the probability c_{j_1} / n. The i-th number c_{j_i} is sampled from the remaining ones with the probability

$$c_{j_i} / (n - \sum_{\nu}^{i-1} c_{j_\nu}), \quad i = 1, 2, \ldots, k-1.$$

Then $(c_{j_1}, \ldots, c_{j_k})$ from $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^u)$ is $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^o)$. The procedure is called "size biased random permutation", which appeared as a heap problem of the computer file organization.

2.3 Residual allocation model

The probabilities (3) of $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^o)$ is expressed as follows

$$p((c_1, \ldots, c_k); \theta, \alpha) =$$
$$\left(\prod_{j=1}^{k-1} \binom{\sum_{i=j}^{k} c_i - 1}{c_j - 1} \frac{(\theta + j\alpha)^{[\sum_{i=j+1}^{k} c_i]} (1-\alpha)^{[c_j - 1]}}{(\theta + (j-1)\alpha + 1)^{[\sum_{i=j}^{k} c_i - 1]}} \right) \frac{(1-\alpha)^{[c_k - 1]}}{(\theta + k\alpha + 1)^{[c_k - 1]}}.$$

At the j-th stage, if the remaining number (of balls) is $r_j = n - \sum_{\nu=1}^{j-1} c_\nu$, $c_j - 1$ follows the negative hypergeometric distribution NgHg $(r_j - 1, 1 - \alpha, \theta + j\alpha)$.

NgHg $(n; \alpha, \beta)$ means the mixture of binomial distribution Bn (n, p) when p is a random variable following the beta distribution Be (α, β). This fact is used to generate the random partition $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^u)$ using quasi random numbers.

The Markovian process generating $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^o)$, that is transform of $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^u)$ to $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^o)$ is a special case of *residual allocation model* developed in the ecology for habitats. If $x \sim$ NgHg $(n; \alpha, \beta)$, x/n converges to Be $(\alpha, \beta), n \to \infty$, and $(c_1/n, \ldots, c_k/n)$ converges to a residual allocation of the interval (0, 1).

Let the standard simplex of countable dimension be denoted by

$$\Delta_\infty = \{x = (x_1, x_2, \ldots);\ x_k \geq 0,\ \sum_{k=1}^{\infty} x_k = 1\}. \tag{5}$$

A probability distribution on Δ_∞ is a random partition of $(0, 1)$.

Let $(W_i)_{i=1}^{\infty}$ be a sequence of independent random variables on the interval $(0, 1)$. Another sequence $(V_i)_{i=1}^{\infty}$ is constructed from it as follows:

$$V_1 = W_1\ ;\ V_2 = (1 - W_1)\, W_2,\ 1 - V_1 - V_2 = (1 - W_1)\,(1 - W_2)\,;$$

and generally

$$\begin{cases} V_k = \left(\prod_{i=1}^{k-1} (1 - W_i)\right) W_k = \left(1 - \sum_{i=1}^{k-1} V_i\right) W_k \\ 1 - \sum_{i=1}^{k} V_i = \prod_{i=1}^{k} (1 - W_i), \quad k = 2, 3, \ldots. \end{cases} \tag{6}$$

The sequence $(V_i)_{i=1}^{\infty}$ is a random variable on Δ_∞, and Markovian, if $\sum_{i=1}^{k-1} V_i = v$, V_k is a random variable on $(0, 1-v)$. The random mechanism generating $(V_i)_{i=1}^{\infty}$ from $(W_i)_{i=1}^{\infty}$ is the residual allocation model. The above discussion is its finite and discrete version.

Proposition 4. *The limit distribution of the ratio of a random partition of* $(c_1/n, \ldots, c_k/n$ *following* $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^o)$ *is the residual allocation model* $(V_i)_{i=1}^{\infty}$, *generated from* $(W_i)_{i=1}^{\infty}$, $W_i \sim \mathsf{Be}\,(1-\alpha, \theta + i\,\alpha)$.

This is a generalization of *GEM* (generalized Engen-McClosley) distribution, which is the case $\alpha = 0$ and $W_i)_{i=1}^{\infty}$ is i.i.d. The limit distribution of the proposition will be denoted by $\mathcal{P}(\theta, \alpha; \Delta_\infty)$.

Conversely, a 'multinomial sample' from $\mathcal{P}(\theta, \alpha; \Delta_\infty)$ is $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^u)$, and if the sampling is sequential and the observation is numbered, the other types are obtained.

3 The number of subsets

The number K_n of subsets of a partition, or the number of nonempty urns plays naturally an important role in random partitions. For example, if $\alpha = 0$, K_n is the sufficient statistic of $\mathcal{P}(\theta, 0; \mathcal{C}_n^u)$, and of all types of random partitions.

In the random partition $\mathcal{P}(\theta, \alpha; \mathcal{A}_n^o)$, $\mathcal{P}(\theta, \alpha; \mathcal{A}_n^u)$, $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^o)$, $\mathcal{P}(\theta, \alpha; \mathcal{C}_n^u)$, the number of subsets K_n $(= \sum_{j=1}^{n} S_j$ is terms of the size index) has the following distribution

$$P\{K_n = k\} = \frac{\theta(\theta+\alpha)\cdots(\theta+(k-1)\alpha)}{\theta^{[n]_\alpha k}} \left|c(n,k;\alpha)\right|, \quad k = 1, \ldots, n, \tag{7}$$

where $|c(n,k;\alpha)| = (-1)^{n-k} C(n,k;\alpha)$, and $C(n,k;\alpha)$ is defined by the following two-variable polynomial identity.

$$(st)^{(n)} = \sum_{k=1}^{n} C(n,k;s) t^{(k)}. \tag{8}$$

$C(n,k;s)$ was named *C-number* by Charalambides (1998), actually it is a polynomial in s of order $n-k$. The factor $|c(n,k;\alpha)| \,/\, \alpha^k$ in the above expression is a polynomial of α of degree $n-k$ with integer coefficients.

Using the stirling numbers of the first kind (unsigned) $\left[{n \atop k}\right]$ and the second kind $\left\{{n \atop k}\right\}$,

$$P\{K_n = k\} = \frac{\prod_{j=1}^{k}(\theta+(j-1)\alpha)}{\theta^{(n)}} \sum_{r=k}^{n} \left[{n \atop r}\right] \left\{{r \atop k}\right\} (-\alpha)^{r-k}, \tag{9}$$

$$s^{[n]} = \sum_{k=1}^{n} \left[{n \atop k}\right] s^k, \quad s^n = \sum_{k=1}^{n} \left\{{n \atop k}\right\} s^{(k)}.$$

4 Limit distribution

The factorial moments of K_n is

$$E(K_n^{(r)}) = \frac{\prod_{i=1}^{r}(\theta+(i-1)\alpha)}{\alpha^r \theta^{[n]}} \sum_{k=0}^{r}(-1)^{r-k}\binom{r}{k}(\theta+k\alpha)^{[n]} \tag{10}$$

$$\sim n^{r\alpha}\frac{\prod_{i=1}^{r}(\theta+(i-1)\alpha)}{\alpha^r \theta^{[n]}}\frac{\Gamma(\theta)}{\Gamma(\theta+r\alpha)}, \quad n\to\infty,$$

hence

$$\mu_r' := \lim_{n\to\infty} E\left[\left(\frac{K_n}{n^\alpha}\right)^r\right] = \frac{(\prod_{i=1}^{r}(\theta+(i-1)\alpha)}{\alpha^r}\frac{\Gamma(\theta)}{\Gamma(\theta+r\alpha)}. \tag{11}$$

This is the r-th moment about origin of the probability density

$$\frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)}x^{\theta/\alpha}g_\alpha(x). \tag{12}$$

where $g_\alpha(x)$ is the probability density of *Mittag-Leffler distribution*, which is characterized by its moment

$$\int_0^\infty x^p g_\alpha(x)dx = \frac{\Gamma(p+1)}{\Gamma(p\alpha+1)}, \quad \forall p > -1.$$

The distribution with the density (12) is called *Generalized Mittag-Leffler distribution* and denoted by GMtLf (θ,α). It is shown from (11) that $K_n/n^\alpha \xrightarrow{\mathrm{d}}$ GMtLf $(n\to\infty)$.

Let $S = (S_1,\ldots,S_n)$ be a size index of $\mathcal{P}(\theta,\alpha;\mathcal{C}_n^u)$, its factorial moments are given by

$$E\left(\prod_{j=1}^{n} S_j^{(r_j)}\right) = \frac{n^{(s)}(\theta+r\alpha)^{[n-s]}}{\theta^{[n]}}\prod_{i=1}^{r}(\theta+(i-1)\alpha)\prod_{j=1}^{n}\left(\frac{(1-\alpha)^{[j-1]}}{j!}\right)^{r_j}, \tag{13}$$

$$r_j = 0,1,\ldots; \quad r := \sum_{j=1}^{n} r_j, \; s := \sum_{j=1}^{n} j r_j.$$

Specially

$$E\left(\frac{jS_j}{n}\right) = \binom{n-1}{j-1}\frac{(\theta+\alpha)^{[n-j]}(1-\alpha)^{[j-1]}}{(\theta+1)^{[n-1]}}. \tag{14}$$

The right hand side is the probability function of NgHg $(n-1;1-\alpha,\theta+\alpha)$, which is the distribution of C_1 of $\mathcal{P}(\theta,\alpha;\mathcal{C}_n^o)$.

From the above moments

$$E\left(\prod_{j=1}^{n}\left(\frac{j!S_j}{\alpha(1-\alpha)^{[j-1]}n^\alpha}\right)^{r_j}\right) \sim \mu_r'. \tag{15}$$

Hence, for any S_j, $j < \infty$,

$$\frac{j!S_j}{\alpha(1-\alpha)^{[j-1]}n^{\alpha}} \xrightarrow{\mathrm{d}} \mathsf{GMtLf}\ (\theta,\alpha), \quad n \to \infty.$$

This means that

$$\left(\frac{1}{\alpha}\frac{S_1}{n^{\alpha}}, \frac{2!}{\alpha(1-\alpha)}\frac{2!\,S_2}{n^{\alpha}}, \frac{3!}{\alpha(1-\alpha)^{[2]}}\frac{S_3}{n^{\alpha}}, \cdots\right)$$

degenerates as $n \to \infty$ to the one-dimensional distribution $\mathsf{GMtLf}\ (\theta,\alpha)$. The sharing proportions

$$\alpha(1-\alpha)^{[x-1]}/x!, \quad x = 1,2,\ldots, \quad 0 < \alpha < 1$$

is a distribution named *Sibuya's distribution*, Devroye(1993).

5 Parameter estimation

The log likelihood of $\mathcal{P}(\theta,\alpha;\mathcal{C}_n^u)$, in terms of size index $s = (s_1,\ldots,s_n)$ is

$$L = \mathsf{const} + \sum_{i=1}^{k-1}\log(\theta+i\alpha) - \sum_{j=1}^{n-1}\log(\theta+j) + \sum_{j=2}^{n} s_j\left(\sum_{i=1}^{j-1}\log(i-\alpha)\right),$$

and the likelihood equation is

$$\begin{aligned}
\frac{\partial L}{\partial\theta} &= \frac{1}{\theta} + \cdots + \frac{1}{\theta+(k-1)\alpha} - \left(\frac{1}{\theta} + \cdots + \frac{1}{\theta+n-1}\right) = 0,\\
\frac{\partial L}{\partial\alpha} &= \frac{1}{\theta+\alpha} + \cdots \frac{k-1}{\theta+(k-1)\alpha} - \sum_{j=2}^{n} s_j\left(\frac{1}{1-\alpha} + \cdots + \frac{1}{j-1-\alpha}\right) = 0.
\end{aligned} \tag{16}$$

Proposition 5. *Let $I(\alpha,\theta)$ be the Fisher information matrix, if $\alpha > 0$, and $n \to \infty$ its elements are*

$$I_{\theta\theta} = O(1), \quad I_{\theta\alpha} = O(\log n), \quad I_{\alpha\alpha} = O(n^{\alpha}).$$

Further, if $\alpha / \log n \to 0$,

$$I_{\theta\theta} = O(\log n), \quad I_{\theta\alpha} = O((\log n)^2), \quad I_{\alpha\alpha} = O((\log n)^3).$$

If $\alpha = 0$, K_n is a sufficient statistics its distribution is in the exponential family (power series family), and its information is $O(\log n)$. In the case $\alpha > 0$ the estimation problem is irregular and difficult. The ML estimator does not look optimal.

Other possible estimators are as follows. Let

$$R_i := \sum_{j=1}^{n} \frac{j^{(i)}S_j}{n^{(i)}}, \quad i = 0,1,\ldots, \quad (R_0 = K_n,\ R_1 = n).$$

The simplest estimation is

$$\hat{\alpha} = S_1 / R_0, \quad \hat{\theta} = (1 - \hat{\alpha}) / R_2 - 1.$$

Using R_2 and R_3, the estimators are

$$\hat{\alpha} = \frac{R_3/R_2 - 2R_2 + R_3}{R_3/R_2 - R_2}, \quad \hat{\theta} = \frac{1 + R_2 - 2R_3/R_2}{R_3/R_2 - R_2}, \tag{17}$$

or solve the nonlinear equation

$$E(R_2) = R_2 \quad \text{and} \quad E(R_0) = R_0 \quad (\text{or} \quad E(S_1) = S_1).$$

Least sum of squares (or other distance) estimator looks favorable:

$$(\hat{\theta}, \hat{\alpha}) = \arg\min \| j\, S_j / n - E(j\, S_j / n) \|.$$

6 Some geneses

$\mathcal{P}(1, 0; \mathcal{A}_n^o)$, $\mathcal{P}(1, 0; \mathcal{C}_n^o)$

A1 A permutation $(\sigma_1, \ldots, \sigma_n)$ of $\mathcal{N}_n$ is a bijection $\mathcal{N}_n \to \mathcal{N}_n$ and the mapping divides $\mathcal{N}_n$ into cycles. If the probabilities of all the permutations are equal to $1 / n!$. The random partitions by cycles is $\mathcal{P}(1, 0; \mathcal{A}_n^o)$.

A2 Let $(X_1, \ldots, X_n)$ be a random sample from a continuous distribution function. $X_{(1)} = X_1$ is the initial maximum record. If $X_1, \ldots, X_{k-1} \le X_{(j-1)}$ and $X_{(j-1)} < X_k$ then $X_{(j)} = X_k$ is the j-th maximum record, $j = 1, 2, \ldots$. A new record breaks the sample into before the record and after it including itself. The numbers of components broken by records is $\mathcal{P}(1, 0; \mathcal{C}_n^u)$.

A3 Let $(X_1, \ldots, X_n)$ be independent and $X_j \sim \mathsf{N}(\mu_j, \sigma^2)$. The maximum likelihood or least squares estimate of $(\mu_1, \ldots, \mu_n)$ under the order restriction $\mu_1 \le \cdots \le \mu_n$ is obtained by Pool Adjacent Violators Algorithm. It is to divided $\mathcal{N}_n$ into the intervals $(j_1 = 1, \ldots, j_2 - 1), (j_2, \ldots, j_3 - 1), \ldots, (j_k, \ldots, j_{k+1} = n)$ such that

$$\hat{\mu}_{j_i} = \cdots = \hat{\mu}_{j_{i+1}-1} = \frac{1}{j_{i+1} - j_i} \sum_{\nu = j_i}^{j_{i+1}-1} X_\nu, \quad \hat{\mu}_1 \le \cdots \le \hat{\mu}_n.$$

Under the null hypothesis $\mu_1 = \cdots = \mu_n$, $(j_2 - j_1, \ldots, n - j_k)$ is the random partition $\mathcal{P}(1, 0; \mathcal{C}_n^u)$.

A4 There are n particles on a line. A particle has mass m_j, is located at x_j, and moving with the velocity v_j, $j = 1, \ldots, n$. If two particles j and k collide they reduce to one particle (or a cluster of particles) with mass $m_j + m_k$, velocity $(m_j v_j + m_k v_k) / (m_j + m_k)$, that is, the completely inelastic collision. After some finite time interval, the particles do not collide, the velocities of particles are ordered according to their positions. The size of final clusters, the number of particles collided into one, is $\mathcal{P}(1, 0; \mathcal{C}_n^u)$ only if the initial velocities $(v_1, \ldots, v_n)$ is an i.i.d. sequence of random variables. Sibuya, et al.(1990).

$\mathcal{P}(\theta, 0; \mathcal{A}_n^o)$

A5 Taga and Isii model of spreading rumor. See, Taga and Isii (1959) and Bartholomew (1967). From an information source I_0, a news, rumor, technology, or knowledge is spread to people of a community in time as a Poisson process with the intensity λ. From the primary informant $I_1, I_2, \ldots$, the news is spread to other people, and from them to others, with the same intensity μ for each. Let a person, excluding I_0, be denoted by B_i, $i = 1, 2, \ldots$ in the order of time when the news is received. When n people, including I_i, have received the news, they are divided into the branches of the root. The group a_i, started from I_i, $i = 1, 2, \ldots$ is a random partition $\mathcal{P}(\theta, 0; \mathcal{A}_n^o)$, $\theta = \lambda / \mu$.

$\mathcal{P}(\theta, 0; \mathcal{C}_n^u)$

A6 The classification of homonyms by the accent, Sibuya(1991)

$\mathcal{P}(\theta, \alpha; \mathcal{C}_n^u)$

A7 Stochastic abundance, Engen (1978), and Statistical disclosure control, Hoshino(2001).

References

[1] Bartholomew, D.J. (1967) *Stochastic Models for Social Processes*, (3rd ed. 1982), Wiley, New York, N.Y.

[2] Charalambides, Ch. A. and Singh, J. (1988) A review of the Stirling numbers, their generalizations and statistical applications, *Commun. Statist.-Theory Meth.* **17**, 2533–2595.

[3] Devroye, L. (1993) A triptych of discrete distributions related to the stable law, *Statist. Probab. Letters*, **18**, 349–351.

[4] Engen, S. (1978). *Stochastic Abundance Models with Emphasis on Biological Communities and Species Diversity*, Chapman and Hall.

[5] Ewens, W.J. (1990). Population genetics theory – the past and the future, Lessard, S. ed., *Mathematical and Statistical Developments of Evolutionary Theory*, NATO Adv. Sci. Inst. Ser. C-299, Kluwer, Dordrecht, 177–227.

[6] Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment, *Journal of Official Statistics,* (to appear).

[7] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997) *Discrete Multivariate Distributions*, Wiley.

[8] Pitman, J. (1995) Exchangeable and partially exchangeable random partition, *Prob. Theory Relat. Fields* **102**, 145–158.

[9] Pitman, J. (1996a) Random discrete distributions invariant under size-biased permutation, *Adv. Appl. Prob.* **28**, 525–539.

[10] Pitman, J. (1996b) Some developments of the Blackwell-MacQueen urn scheme, in T. S. Ferguson, L. S. Shapley and J. B. MacQueen (eds.), *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell* (IMS, Haywars, CA) pp. 245–267.

[11] Pitman, J. (1997) Partition structure derived from Brownian motion and stable subordinaters, *Bernoulli* **3**, 79–96.

[12] Pitman, J. and M. Yor (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator, *Ann. Probab.* **25**, 855–900.

[13] Pitman, J. (1999) Characterizations of Brownian motion, bridge, meander and excursion by sampling at independent uniform times, *Electronic J. Probability* **4**, Paper 11, 1–33.

[14] Sibuya, M. (1991) A cluster-number distribution and its application to the analysis of homonyms, *Japanese J. Appl. Statist.* **20**, 139-153. (in Japanese)

[15] Sibuya, M., Kawai, T., and Shida, K. (1990) Equipartition of particles forming clusters by inelastic collisions, *Physica* **A 167**, 676-689.

[16] Taga, Y. and Isii, K. (1959), On a stochastic model concerning the pattern of communication -Diffusion of news in a social group-, *Ann. Inst. Statist. Math.*, **11**, 25–43.

[17] Yamato, H. and Sibuya, M. (2000) Moments of some statisitcs of Pitman Sampling formula, *Bulletin of Informatics and Cybernetics*, Fukuoka, **32**, 1–10.

[18] Yamato, H., Sibuya, M. and Nomachi, T. (2001) Ordered sample from two-parameter GEM distribution, Statistics and Probability Letters,(in print)